La publication durable digitale des guides d'archives de l'histoire du 20ème siècle

Veerle Vanden Daelen, Jennifer Edmond, Petra Links, Mike Priddy, Linda Reijnhoudt, Václav Tollar, Annelies van Nispen, Charlotte Hauwaert, Charles Riondet

La publication durable digitale des guides d'archives de l'histoire du 20ème siècle

Auteurs : Veerle Vanden Daelen, Jennifer Edmond, Petra Links, Mike Priddy, Linda Reijnhoudt, Václav Tollar, Annelies Van Nispen

Traduction : Charlotte Hauwaert (CegeSoma)

Rédaction : Charles Riondet (The French Institute for Research in Computer Science and Automation, INRIA) et Veerle Vanden Daelen (Kazerne Dossin : Mémorial, Musée et Centre de Documentation sur l'Holocauste et les Droits de l'Homme)

Avant-propos (Veerle Vanden Daelen)

Depuis le début de ma participation au projet *European Holocaust Research Infrastructure* (EHRI) en 2011, j'ai travaillé avec une multitude d'institutions conservant des archives historiques du 20e siècle allant des centres d'archives classiques aux lieux de mémoire, bibliothèques et archives privées. L'intégration dans un environnement unique des descriptions archivistiques– quand celles-ci étaient disponibles – relatives aux sources de la Shoah conservées par ces différentes institutions s'est avérée un défi majeur. Tout au long de cette aventure, mes collègues et moi-même avons fait connaissance avec des personnes – des archivistes professionnels mais aussi des volontaires – dévouées qui préservent et décrivent ces sources, parfois en dépit de contraintes infrastructurelles, d'équipe, ou de temps disponible. Aucune institution patrimoniale sollicitée par EHRI n'a su exporter les descriptions d'archives sans un traitement préalable plus ou moins approfondi. Ces mêmes expériences furent observées par le projet CENDARI (Collaborative EuropeaN Digital Archival Research Infrastructure). Par conséquent, nous avons soumis une proposition pour l'appel Open Humanities 2015, « Open History : Sustainable digital publishing of archival catalogues of twentieth-century history archives » de DARIAH. Cette proposition fut particulièrement portée par: Jennifer Edmond (Trinity college Dublin), Petra Links (en 2015 NIOD Institute for War, Holocaust and Genocide Studies, Amsterdam, maintenant Vrije Universiteit Amsterdam), Mike Priddy (Data Archiving and Networked Services, DANS, La Haye), Linda Reijnhoudt (DANS, La Haye), Václav Tollar (Institute for the Study of Totalitarian Regimes and Security Services Archive, USTR and ABS, Prague), Annelies van Nispen (NIOD, Amsterdam), et moi-même (en 2015 Centre d'Études et Documentation Guerre et Sociétés contemporaines – CEGESOMA Bruxelles, maintenant Kazerne Dossin : Mémorial, Musée et Centre de Documentation sur l'Holocauste et les Droits de l'Homme).

Le but principal du projet "Open history" était d'enrichir le dialogue entre les fournisseurs de (méta)données et les infrastructures de recherche. Dans ce but nous avons développé deux outils : un article accessible et générique concernant l'importance des pratiques de normalisation et comment

atteindre ce but ; et un modèle qui fournit des listes de contrôle pour une auto-évaluation des institutions archivistiques. L'article que nous avons écrit (reproduit ci-après) reste volontairement un article générique, sans jargon, pour que les lecteurs non archivistes ou informaticiens puissent facilement accéder aux informations données. Nous espérons dès lors que le texte sera facilement compréhensible aussi bien pour les personnes qui décrivent les sources dans les archives (avec ou sans qualifications informatiques et/ou archivistiques) que pour les responsables (directeurs et conseils consultatifs) afin de comprendre les avantages d'investir dans la normalisation et le partage de données.

Il est important de remarquer que ce texte est une première étape, et non un résultat figé. Le texte n'aborde pas tous les aspects de la normalisation et de la publication de (méta)données, ni les mises à jour ou retours pour annotations et commentaires. L'idée est de pouvoir utiliser le texte en son entièreté ou certaines parties, et que d'autres chapitres et sections s'ajouteront au fil du temps et grâce à la participation d'autres communautés qui s'empareront du texte. Certaines institutions liront cet article et y verront une validation de leur méthode, d'autres – particulièrement les petites institutions, comme les institutions mémorielles privées – retrouveront dans ce texte une introduction facile et une approche pragmatique pour les aider dans leurs efforts.

Contenu

Introduction

En règle générale, quand des représentants d'une infrastructure de recherche contactent des centres d'archives, ils sont intéressés par les données et/ou métadonnées numériques disponibles et par l'acquisition (semi)automatique de ces (méta)données dans un format spécifique (dépendant d'un logiciel, en concordance avec les directives du projets, etc.). Le point de discussion principal est la façon dont les informations relatives aux fonds et aux collections peuvent être publiées et partagées avec des chercheurs et le grand public. Souvent, ces questions ne sont pas faciles à résoudre et demandent un contact suivi avec plusieurs membres de l'institution. Cette démarche peut entraîner un doute de l'institution envers le projet, les conservateurs pouvant se demander pourquoi ils devraient investir du temps et de l'énergie pour le bénéfice d'une équipe de recherche, d'autant plus quand ces efforts vont à l'encontre de son propre savoir-faire. Il y a plusieurs bonnes raisons pour lesquelles s'engager dans une infrastructure englobante est un avantage clé pour votre institution, le premier étant la mise en valeur et la visibilité internationale de votre institution.

Avant de pouvoir partager toute information de manière efficace, il faut procéder à une normalisation de l'information. La standardisation est non seulement un prérequis pour partager les données de manière efficiente et durable, mais elle apporte beaucoup d'avantages pour votre institution car elle pousse votre institution à générer un processus de travail explicite et documenté. La connaissance du savoir-faire concernant vos fonds et collections ne sera donc plus limitée à une ou deux personnes, mais cette information sera disponible pour tout le monde.

La nécessité de la normalisation dans un monde numérique : Les humains parviennent à vivre avec des données de (très) faible qualité, les ordinateurs non.

Au cours des dernières décennies, les ordinateurs et les outils numériques ont considérablement transformé les méthodes d'accès et de traitement des données. Il est certes toujours possible de chercher dans les archives au moyen de fiches cartonnées et d'instruments de recherche imprimés, toujours utiles pour un premier contact avec les données conservées. Pour autant, ces méthodes ont progressivement été numérisées. Des pionniers, et peut-être que vous en faites partie, ont investi beaucoup de temps et d'efforts dans le développement de solutions sur mesure pour les besoins spécifiques et les contraintes de leurs institutions, bien avant que des outils (gratuits et ouverts) soient accessibles. Il se peut que votre institution ait cherché de l'aide de l'extérieur en acquérant un logiciel à un éditeur. Il est possible aussi que vous ayez publié vos guides de recherche sur votre site web, pour vous rendre compte ensuite que lorsque des infrastructures de recherche vous sollicitent, la transmission de vos données aux chercheurs est difficile voire impossible.

Certes, ces données et outils sont compréhensibles pour les lecteurs humains, mais les ordinateurs qui doivent intégrer toutes les informations ne parviennent ni à les reconnaître ni à les traiter : ce qui est décrit par votre institution, avec ses propres spécificités, ne correspond pas nécessairement aux méthodes de descriptions d'une autre institution. Par exemple, si vous trouvez dans un guide de recherche que certains documents sont en « nérlandais et en françai », vous comprenez parfaitement le

sens sémantique malgré les fautes d'orthographe. Si des abréviations sont adoptées en interne, vous pouvez identifier que les langues sont le néerlandais et le français par les abréviations « N » et « F ». En revanche, les personnes qui ne sont pas au courant des codes internes pour les langues, auront des soucis pour les déchiffrer, en raison de leur ambiguïté : cela peut aussi bien être le norvégien ou le finlandais par exemple. Un ordinateur, en revanche, ne peut déchiffrer le sens sémantique dans les deux cas puisqu'il se base sur des formats standardisés (comme par exemple des codes de langues « NL » et « FR ») qui lui permettent d'organiser l'information provenant de sources différentes. En gros, les ordinateurs nous poussent à normaliser notre méthode de travail, tout comme les infrastructures de recherche qui se présentent dans votre institution. C'est une bonne chose ! Premièrement parce que ceci permet aux infrastructures d'atteindre leur objectif et deuxièmement parce que réfléchir à une normalisation de vos méthodes de travail dans votre institution et la documenter vous offriront non seulement une connaissance et une compréhension plus complète de vos propres fonds (une recherche portant sur tous les documents avec le code de langue standard « FR » permet de ne pas rater les documents en « françai »). Mais cela offre aussi d'autres moyens de préservation et de présenter vos ressources: les documents en « néerlandais et français » ou « Dutch and French » peuvent être présentés aux utilisateurs de différentes manières en fonction de leurs besoins. Si nous croyons en notre mission de préserver notre héritage culturel, tant physique que numérique, nous devons nous adapter et reconnaître les avantages que ces nouvelles méthodes et nouveaux outils nous offrent.

Pas de panique – personne n'est parfait

Vous pouvez penser que vous êtes loin du compte, mais il est important de prendre conscience qu'en général, la plupart des institutions sont toujours à la recherche de méthodes et de bonnes pratiques applicables à leur institution, adaptées non seulement à leur contenu scientifique, mais également à leurs moyens humains et financiers. Des centres d'archives qui semblent parfaits vu de l'extérieur laissent pourtant souvent des points à perfectionner si nous regardons de plus près. Être une institution avec une longue histoire n'est pas nécessairement un avantage, car bien souvent, les systèmes d'informations se sont empilés sans se remplacer, engendrant une structure unique mais complexe qui peut s'avérer difficile à démêler. Héritant de connaissances accumulées et de pratiques entremêlées, ces institutions peuvent être confrontées à des difficultés pour s'ajuster aux opportunités offertes par le numérique. Par conséquent, il n'y a pas de solution toute faite ou d'approche universelle qui permettrait à tous les centres d'archives de standardiser la description de leurs collections. De la même manière, toutes ces institutions n'ont pas les mêmes besoins, en termes de complexité, pour leurs outils et méthodes. Cependant, toutes souhaitent informer le monde extérieur de leur existence et de leurs activités.

Ce rapport détaille les expériences de deux projets d'infrastructure de recherche pan-européens qui ont tenté de rassembler des données sur des institutions conservant des archives contemporaines (XX^{e} siècle) et leurs collections (au moins au niveau du fonds ou de la série) dans des environnements virtuels de recherche. Ces expériences sont déterminantes pour évaluer la situation actuelle au sein de ces institutions, et savoir dans quelles mesures elles sont prêtes à se lancer dans de tels projets est pertinent pour les infrastructures de recherches elles-mêmes. Par ailleurs, la restitution des leçons tirées

de ces projets aux institutions qui détiennent les collections peut être également être considéré comme une avancée à moyen ou long terme.

Que font les infrastructures de recherche comme CENDARI et EHRI ?

Les deux projets, qui ont partagé leurs expériences et connaissances pour écrire cet article, sont CENDARI et EHRI. Les acronymes signifient : *Collaborative EuropeaN Digital Archival Research Infrastructure* et *European Holocaust Research Infrastructure*. Les deux projets sont financés par la Commission Européenne et font partie du programme *Research and Innovation*. Tous deux ont pour matériel de base des archives historiques du vingtième siècle. Pour pouvoir développer un environnement virtuel de recherche (EVR), les projets souhaitaient rassembler les informations sur des collections dispersées dans un grand nombre de pays et d'institutions, dans le but de : 1. Mettre à disposition un point d'accès centralisé pour l'information et 2. Fournir des méthodologies et outils pour la recherche qui seraient inaccessibles autrement. Par conséquent, le but n'est pas de rendre les institutions d'archives superflues, mais au contraire de rassembler leurs informations dans un portail, ce qui augmente la visibilité et les connaissances sur les institutions participantes et leurs collections. Les projets des infrastructures de recherche ne doivent pas être perçus comme une tentative de s'approprier l'héritage culturel des institutions, mais plutôt comme des utilisateurs performants qui requièrent des besoins spécifiques à grande échelle. Par ailleurs, en participant, les institutions deviennent parties intégrantes de communautés centrées autour du même type de sources (Moyen Âge et Première guerre mondiale pour Cendari, la Shoah pour EHRI).

Comment ces infrastructures ont-elles planifié l'intégration de (méta)données dans ces portails ? CENDARI et EHRI ont rassemblé un ensemble de compétences, historiens, archivistes, spécialistes des Humanités numériques et informaticiens, venues du monde entier. Le processus implique toutes les disciplines, avec le but de garantir que le contenu correct parvienne dans le bon format. Par contenu correct, nous entendons ce qui est dans le périmètre du sujet de recherche (Moyen Âge et Première Guerre Mondiale pour CENDARI, la Shoah pour EHRI), le bon format fait référence à la structure des données (les standards, voir ci-dessous). En théorie, le travail de ces équipes interdisciplinaires semble être simple: contacter toutes les archives, recevoir leurs métadonnées (y compris les données décrivant l'institution et ses collections) et les incorporer dans le système. La pratique s'avéra néanmoins différente car les métadonnées n'étaient souvent pas disponibles. Et si elles l'étaient, elles n'étaient pas standardisées ou disponibles dans un format qui puisse être utilisé en dehors de l'institution même. Il a donc fallu réorganiser et nettoyer les métadonnées pour qu'elles puissent être partageables. Ceci ne veut pas pour autant dire que toutes les métadonnées devenaient complètement uniformes, mais plutôt que les métadonnées que les institutions étaient prêtes à partager avaient besoin d'être synchronisées vers un dénominateur commun pour que l'importation dans le système puisse avoir lieu.

Tout le monde est différent (et les solutions sur mesure sont coûteuses et rarement durables)

Arriver à ce but a nécessité du changement et donc un travail approfondi. Comme nous le savons tous, aucun centre d'archives – à notre connaissance – n'a à sa disposition une « équipe de réserve » qui n'attend que d'être sollicitée pour accomplir sa mission, et aucun ne possède un budget illimité pour recruter de tels équipes. En réalité, la plupart des institutions qui préserve notre héritage culturel et les documents historiques sont sous-financées et en sous-effectifs. Leur chance est d'avoir des employés motivés qui font tout leur possible pour préserver les archives et pour les rendre disponible aux chercheurs. Qu'ont fait CENDARI et EHRI, dans ce contexte, pour intégrer les données dans leurs portails respectifs? Le site de CENDARI expose que « l'intégration des données des institutions, aux niveaux de la collection et de la pièce, requiert un processus personnalisé en utilisant des outils XML et des schémas sur mesure. » L'emploi des mots « personnalisé « et « sur mesure » indique qu'il n'existe pas de méthode universelle applicable à toute situation. Les deux projets avaient donc intérêts à exceller dans des approches sur mesure. CENDARI utilise notamment des outils qui permettent de moissonner le contenu de pages web (web scraping), permettant d'importer l'information en tant que telle. Au sein d'EHRI, les méthodes de travail vont de l'ajout manuel des descriptions à l'importation massif. Toutefois, la majorité des cas se situaient dans un entre deux et nécessitaient le recours à une mise en correspondance complète et un traitement préalable des données avant qu'elles puissent être intégrées au portail en ligne. Ceci n'est ni compliqué ni extraordinaire, à condition que les méthodes utilisées puissent être applicables d'une institution à l'autre. En 2011, la première année du projet EHRI, la boutade « chaque archive est unique » fut récitée en riant. À la fin de la première phase de EHRI, quatre ans plus tard, les équipes chargées de l'implantation des données riaient moins, ayant constaté que le nombre d'approches sur mesure équivalait presque le nombre d'institutions intégrées. La phrase « chaque archive est unique » est en fait très vraie : chaque institution a sa propre histoire, ses collections spécifiques et sa mission, et par extension sa propre approche pour la description et la manipulation des collections. La même frustration entraîna CENDARI à façonner un organigramme pour envisager toutes les façons de travailler avec les institutions, avec l'espoir d'assurer ainsi que toutes les potentielles réutilisations de méthodes et d'outils pourraient être facilement identifiées.

Malgré la grande variété au sein des archives, CENDARI et EHRI ont su intégrer du contenu dans leurs portails respectifs. Mieux, il a ainsi été possible de révéler beaucoup « d'archives cachées » à la communauté des chercheurs. Cette implémentation n'est toutefois, malheureusement, pas encore durable. Idéalement, l'incorporation des données devrait être un processus répétable, sans devoir ajuster les méthodes de travail à chaque fois. Les institutions et les projets devraient pouvoir accéder facilement aux mêmes données (actuellement, les descriptions ne sont pas nécessairement accessibles tant pour l'institution qu'au sein des projets). De ce fait, les mises à jour appliquées à un environnement ne sont pas appliquées dans l'autre environnement.

Par exemple, une institution avec laquelle EHRI travaille, a numérisé toutes ses collections en haute définition, disponibles pour un usage interne via des milliers de scans, mais le site web ne montre que les informations de base sur ces collections. Les milliers de scans ne sont donc pas accessibles dans leur entièreté pour la recherche en dehors de l'institution mais aussi au sein de l'institution même. Ce

processus de numérisation ne propose pas de solution au problème de « l'aiguille dans la botte de foin » car il n'existe toujours pas de descriptions standardisées et facilement accessibles qui peuvent aider à trouver les informations. Si vous recherchez de la correspondance dans des archives privées données à cette institution, le site vous indiquera seulement que des archives privées ont été léguées à l'institution. Il n'y a donc pas de noms ou descriptions détaillés de l'archive privée sur le site de l'institution. De plus, même en salle de lecture, vous n'avez pas d'instruments de recherche. Le chercheur devra donc demander à l'archiviste qui devra se baser sur sa propre expertise pour savoir quels dossiers consulter et trouver les fichiers qui contiennent des informations utiles au chercheur.

Une autre frustration s'ajoute puisque les descriptions sur le site web de l'institution ne sont pas standardisées et qu'il s'agit donc d'un texte plat qui ne peut être intégré en tant que tel dans la base de données de l'infrastructure de recherche. Un facteur de complication supplémentaire est le fait que l'institution n'attribue pas d'identifiants stables et uniques à ses collections : les identifiants changent à chaque fois que le site est mis à jour. Tout ceci crée non seulement des problèmes pour les infrastructures de recherche, mais complique et rend aussi l'organisation des (méta)données dans l'institution chaotique. C'est aussi une des raisons pour lesquelles l'activité des infrastructures de recherche apporte aux institutions une prise de conscience renforcée et un dynamisme pour réfléchir aux prochains éléments, que nous encadrerons dans un modèle *capability maturity*.

Comprendre votre organisation et sa gestion des données par le modèle *capability maturity*

Le « modèle *capability maturity*» (CMM) est tiré d'un système développé dans une étude commanditée par l'armée américaine dans les années 1980. Alors que l'utilisation des ordinateurs se répandait rapidement depuis les années 1960, l'armée a naturellement adopté des systèmes d'information numériques. Ces systèmes se sont développés et ont changé au fil du temps, et de plus en plus de nouveaux systèmes se sont ajoutés et ont été mis en pratique. La somme de tous ces développements a abouti à un système plus efficient et utilisable au mieux, mais aussi à bon nombre d'échecs, et un manque de clarté dans la gestion des systèmes. Le fait d'avoir les outils les plus performant et les logiciels les plus sophistiqués ne garantit pas une performance hors norme ou une meilleure gestion de l'information. Au contraire, l'application de plusieurs programmes, méthodes et outils qui sont insuffisamment intégrés dans l'organisation aboutit à une situation chaotique et des coûts élevés pour les formations et l'amélioration des systèmes. Le CMM a été développé pour aider les organisations à remettre de l'ordre dans le chaos de leurs systèmes d'information. Il documente et structure le comportement d'une organisation, ce qui aide aussi le public à comprendre l'organisation et la gestion des données de l'institution.

Le CMM distingue cinq niveaux de maturité. Le premier niveau – qui est malheureusement le plus fréquent dans les archives historiques du vingtième siècle – est le niveau initial, lorsque les institutions ne documentent pas leur processus de gestion des données. Le système est donc un système « ad hoc » incontrôlable et est soumis à d'éventuels changements. Cet environnement instable de travail ne facilite pas la compréhension du système de gestion des données, même pour les personnes faisant partie de

l'institution. Par exemple : imaginons que Sophie décrit et intègre des descriptions de collections dans la base de données de son institution. Tant qu'il n'existe pas de document détaillant le processus de l'encodage, aucun de ses collègues ne pourra la remplacer ou ne pourra exactement suivre sa procédure. L'institution de Sophie pourrait passer au niveau deux (le niveau répétitif) du modèle CMM si elle mettait au point un document (avec ou sans ses collègues) détaillant sa façon de procéder, ce qui permettrait de répéter les actions et la procédure. Même si le document n'est pas gravé dans le marbre ou n'est pas connu par toutes les personnes à l'institution, il permet d'avoir une trace de documentation des procédures à suivre et permet de les répéter. Si ses collègues aspirent au troisième niveau (le niveau défini), ils auront besoin de plus de définitions poussées et de standardisation. Les méthodes pour décrire les collections de l'institution devront donc intégrer des procédés normalisés, utilisant des normes valables et cohérentes pour toute l'organisation. Les deux prochaines étapes, niveau quatre (gestion) ou niveau cinq (optimisation), font en sorte que les activités acquièrent un niveau plus élevé de maturité et de capacité. Il s'agit respectivement d'une gestion quantitative et de l'application des meilleures pratiques, un niveau qui n'est atteint que par peu d'organisations dans le monde.

L'important, pourtant, est de se rendre compte du besoin crucial d'adopter des méthodes de travail transparentes et reproductibles. Ces méthodes sont nécessaires pour tout travail académique et, pour la recherche historique, devraient idéalement être applicables en amont. De plus, dans les écosystèmes numériques, favoriser la communication est une condition sine qua non. La preuve que des méthodes de travail bien documentées, transparentes et reproductibles étaient absentes dans la plupart des institutions – ce qui signifie que la plupart se trouvent inopportunément au premier niveau, le niveau le plus bas du CMM – est apparue quand CENDARI et EHRI ont voulu obtenir plus d'informations sur les logiciels et les standards archivistiques employés par les institution d'archives pour décrire leurs collections. Même les institutions ayant des outils numériques et utilisant des standards archivistiques, avaient du mal à répondre aux questions, puisque les réponses n'étaient pas disponibles dans un document décrivant la procédure et qu'il fallait donc échanger avec plusieurs personnes de l'institution pour obtenir les informations. Souvent, les personnes qui décrivent les collections n'étaient pas des archivistes de formation, ou bien les bases de données étaient acquises auprès d'entreprises extérieures sans que l'institution ou une personne de l'institution ait la permission (ou la capacité) complète de les gérer. La majorité des institutions ne possède pas de documentation détaillant la méthode de travail qui donne une vue d'ensemble. De plus, s'engager dans une infrastructure entraîne comme conséquence qu'il faut au moins contacter trois personnes : la première pour détailler le « (pour)quoi » (qu'est-ce qui est décrit et est pertinent pour le portail), la suivante pour le « comment » (logiciels, standards et interopérabilité) et la dernière qui dit « où » (l'autorité qui donne la permission d'intégrer les données dans le portail).

Même pour les systèmes les plus « chaotiques » rencontrés par CENDARI et EHRI, une solution fut trouvée. L'incorporation de données d'une institution dans le portail des infrastructures de recherche en employant les méthodes énumérées ci-dessous pour retravailler les données et aboutir à une unique importation dans le portail, est utile pour l'infrastructure, mais a des avantages limités pour l'institution. Il est plus intéressant et plus fructueux si l'institution s'engage dans le dialogue et veut développer sa connaissance de soi et ses capacités. Nous vous invitons donc à continuer à lire pour que vous sachiez ce

qui est possible, quelle aide est à votre disposition, et quels avantages la normalisation et la documentation des pratiques peuvent apporter à votre organisation – chacun avec ses propres capacités, sans devoir repartir à zéro ou devoir investir des sommes exorbitantes dans de nouveaux systèmes. Mike Priddy, Linda Reijnhoudt et leurs collègues au Data Archiving and Networked Services (DANS, La Haie) ont introduit ce CMM dans notre groupe et sont en train de développer ce modèle, tout comme ils maintiennent une liste de contrôle qui introduit des points de référence extérieurs, illustrée avec des exemples pour impliquer votre institution et améliorer vos propres capacités à décrire, publier et partager vos données.

Fournir à votre institution son propre identifiant unique, une lapalissade ?

Le premier pas et le plus évident – mais souvent oublié –, est de décrire votre propre organisation pour pouvoir mettre en avant votre institution. Ceci représente « le plus haut point d'accès » pour votre institution : ce que vous êtes, où vous êtes localisés et quelles collections vous conservez. C'est exactement pour cette première tâche, la description de manière uniforme et aussi complète que possible, que CENDARI et EHRI ont dû investir beaucoup de travail manuel. Dans beaucoup de cas, cela demandait une recherche approfondie sur le site web de l'institution (si celui-ci était disponible) pour collectionner le(s) nom(s) officiel(s), les adresses et les informations de contact, l'histoire, la mission et les informations concernant les collections (comme les références des guides de recherche en lignes ou imprimés). Le résultat est que les autres font leurs propres descriptions de votre institution et qu'une archive nationale, qui est par exemple décrite dans CENDAR et EHRI, sera décrite de plusieurs manières différentes. « Ce ne sont que des variations sur le même thème » pourrait-on dire, mais le problème se manifeste avec les ordinateurs qui ne sont pas nécessairement aptes à identifier ces descriptions comme étant de la même institution. Une institution peut s'auto-identifier comme « State Archives of Belgium», mais EHRI peut utiliser la dénomination « State Archives and Archive in the Provinces » et CENDARI peut choisir de ne pas employer l'anglais et de garder le nom « Archives générales du Royaume et Archives de l'État dans les provinces ». Un ordinateur ne peut donc pas comprendre que les trois appellations se réfèrent à la même entité.

Alors comment résoudre ce problème ? Il n'y a qu'une solution garantie, l'utilisation d'un code d'identification unique et standardisé pour votre institution. Il existe un standard (le code ISIL : « Identifiant international normalisé pour les bibliothèques et les organismes apparentés ») qui, s'il est employé correctement, permet de diffuser un identifiant à quiconque qui veut faire référence à votre institution, tout comme toutes vos descriptions pourront être connectées entre elles sans ambiguïté. Si vous êtes à la recherche de normalisation pour organiser la description de votre institution, un standard existe qui peut vous aider à organiser cette information en suivant un schéma fixe qui rassemble toutes les éléments utiles dans une description utile aussi bien en interne qu'en externe. Cette lapalissade n'est donc pas une si mauvaise idée : elle apporte à votre institution sa propre identité personnelle ce qui permet à l'identité de « voyager » en sécurité et sans ambiguïté à travers le monde numérique.

La prochaine étape est d'inclure d'autres informations de base en plus du code ISIL de votre institution ou « code passeport pour les archives ». Comme dans les passeports humains, il n'y a pas seulement un numéro d'identification – le numéro de registre national ou le numéro de sécurité sociale – mais aussi des champs standardisés comme nom, prénom(s), date de naissance, nationalité, etc. Il existe ainsi un standard qui offre des champs pour décrire votre institution : « International Standard for Describing institutions with Archival Holdings » (ISDIAH), que les archivistes appellent plus naturellement par son abréviation : ISDIAH. Un autre standard fréquemment employé est l'« Encoded Archival Guide » (EAG) qui est par exemple utilisé dans le Portail européen des Archives. Ce projet procure un accès à l'information sur des matériaux archivistiques de différents pays européens et des informations sur les institutions archivistiques à travers le continent.

Une fois que vous avez lu les directives pour les utilisateurs de ces standards (gratuitement sur le web, dans plusieurs langues, voir les liens), vous réalisez que ceux-ci sont des outils très utiles pour organiser l'information d'une institution, pour des éléments basiques comme le nom officiel, l'adresse, les heures d'ouverture et les informations de contact, jusqu'à l'histoire, la gestion et la politique de collecte des documents, une description générale des collections archivistiques et des instruments de recherche, guides et publications. Cela vaut la peine d'analyser les différents champs de ces standards et de travailler avec ceux-ci, non pour les infrastructures de recherche qui souhaitent rassembler cette information, mais pour publier ceux-ci sur votre propre site web. Cela peut paraître une évidence, mais tout centre d'archives devrait pouvoir fournir sur son site internet un exposé sommaire et clair de l'histoire de l'institution et ses mandats, la politique d'acquisition ou des informations claires de contacts. Nous vous encourageons donc à faire ce premier pas facile pour améliorer la visibilité et l'information sur votre institution envers le monde extérieur.

Des classeurs à tiroir et catalogues imprimés jusqu'aux métadonnées numériques

Au-delà de réaliser une description normalisée de votre institution, il est évident que l'essentiel est de décrire les collections de votre institution, les précieuses archives conservées physiquement ou numériquement, aussi bien que les autres types de collections possibles (photos, documents audiovisuels, films, objets, etc.). Il n'y a pas si longtemps, les départements de sciences humaines des universités expliquaient dans des cours de méthodologie l'usage des classeurs à tiroirs et des guides de recherche imprimés. Il y a une vingtaine d'années encore, cette méthode était enseignée aux étudiants, qui apprenaient à constituer un classeur à tiroir – plutôt qu'une bibliothèque digitale – au cours de leur cours académique en Histoire. Si nous parlons aujourd'hui de bases de données et d'outils de gestion des données numériques, nous avons vécu une évolution rapide, marquée par des courbes d'apprentissage accélérées et des changements continus dans les systèmes. Au point que la réponse habituelle lorsqu'il est demandé comment fonctionne le système de gestion des (méta)données employé par une institution est soit que « le système est en cours d'adaptation », soit que « nous sommes en en train de changer de système ». Assez fréquemment, l'enthousiasme créé par de nouvelles tendances génère de la confusion puisque de nouveaux systèmes sont ajoutés à des systèmes

déjà existants, ce qui rend peu clair les descriptions proprement dites, les éventuels chevauchements et autres problèmes.

Le tournant numérique est d'autant plus difficile quand les institutions sont anciennes. Passer des fichiers d'index, catalogues imprimés et instruments de recherche pensés comme des livres, donnant des informations détaillées sur les créateurs et l'histoire de la collection, aux données numériques n'est pas une chose aisée. Le volume de l'information est souvent si vaste que tout transférer vers l'univers numérique ne peut se faire du jour au lendemain. Souvent, le résultat est la coexistence de plusieurs systèmes au sein d'une même institution, avec des chevauchements divers et peu clairs. Bien souvent, le premier pas est de mettre en ligne les « instruments de recherche classiques ». Pour beaucoup d'institutions, cette démarche fut un choc culturel et une transition compliquée : avant l'âge numérique, chaque lecteur du centre d'archive était « connu », ayant travaillé en salle de lecture et/ou contacté un archiviste pour se renseigner sur les collections. En mettant des catalogues et instruments de recherche en ligne, les archives ne peuvent plus contrôler qui consulte leurs données. Un assez grand nombre d'institutions publie leur instrument de recherche en format PDF. Même si ce format est une évolution positive, il ne facilite pas les recherches systématiques et certainement pas si la recherche va au-delà du document même. Ouvrir un grand nombre de PDFs et rechercher systématique dans ceux-ci est également fastidieux. Par conséquent, il est intéressant de connaître les différentes possibilités que les descriptions normalisées, mises à disposition dans des bases de données interopérables, peuvent offrir des possibilités au-delà des publications imprimées ou de leur représentation numérique en ligne.

Pour compliquer davantage les choses, de nouvelles entités intermédiaires sont apparues : des agrégateurs, bibliothèques et centres d'archives numériques, et bien évidemment des infrastructures et des portails de recherche. Certaines d'entre elles, comme Europeana (une plate-forme en ligne donnant accès aux collections numérisées à travers l'Europe), ont pour but de rassembler des contenus numériques préexistants dans un grand portail de recherche commun. D'autres, comme CENDARI et EHRI, se focalisent sur les descriptions des collections, et non les scans ou images numériques des collections, afin de faciliter la recherche pour les chercheurs. Ils aspirent à être des outils de découverte, une sorte de GPS pour les chercheurs permettant d'arriver à leur destination, c'est-à-dire les sources qui les intéressent. En revanche, assembler toutes les pièces pour développer un GPS n'est pas facile, en particulier à cause du fait que les institutions possèdent différents formats et méthodes de descriptions des collections. Comment alors gérer tout cela ? Examinons quelques exemples des meilleures pratiques en commençant par la description des collections.

Comment trouver une aiguille dans un meule de foin : décrire vos collections

Décrire vos collections efficacement est probablement aussi important que la préservation des sources. Sans descriptions précises, trouver le bon document et automatiser le processus de recherche est impossible. Dans les classeurs à tiroir, toutes les fiches respectent la même structure : dans les instruments de recherche, la table des matières rend compte de la structure et toutes les descriptions présentes dans l'instrument de recherche respectent un modèle standard. Des variations entre les

institutions sont possibles, mais les fiches et les instruments de recherche n'avaient pas vocation à être intégrés partiellement ou complètement dans une grande base de données – en tant qu'objets physiques, ils restaient dans un espace physique spécifique et n'étaient « agrégés » que mentalement et à travers l'expérience humaine des utilisateurs. Le numérique offre une alternative, sans la médiation d'une interprétation humaine. Voici le défi principal : comment assurer que toutes ces descriptions des sources et toutes ces métadonnées soient accessibles et interrogeables simultanément, et donc comment élargir au maximum au plus haut degré la disponibilité et la visibilité des information sur les sources réparties dans différents lieux, autant dans le monde physique que numérique ?

Respecter les standards archivistiques

Avant l'âge numérique, les archivistes, et la science archivistique, travaillaient pour procurer un cadre normalisé pour les descriptions des collections. Les standards produits proposaient un modèle normalisé pour composer les descriptions. Ceci ne veut en aucun cas dire que ces standards posent des questions difficiles aux archivistes, mais plutôt qu'ils organisent l'information que les archivistes recueillent dans des champs fixes. Pour les descriptions archivistiques (des collections ou fonds), la « norme générale et internationale de description archivistique » (ISAD(G)) et l'« Encoded Archival Description » (EAD) sont utiles. La lecture de ces standards permet de constater que, aussi intimidants qu'ils puissent paraître, ils suivent une logique très simple. Par exemple, le champ « intitulé » vous invite à renseigner le titre de votre collection ou encore le champ « présentation du contenu » dans lequel vous décrivez le contenu de votre collection. Même si ces standards proposent un grand nombre de champs possibles pour décrire les descriptions des collections d'archives, seulement une partie des champs est obligatoire. Le renseignement des champs non obligatoires dépend de la politique interne de l'institution ou de la nécessité de remplir les champs. Un avantage majeur que les standards procurent est de faciliter l'accessibilité à toutes les descriptions des champs en ligne et pour les standards de du Conseil international des archives (ICA) ; ils sont disponibles en plusieurs langues. Mieux encore, la plupart des bases de données en accès ouvert comme ICA-AtoM (un logiciel qui permet de générer des descriptions d'archives respectant les standards ICA ; Ato M : Access to Memory) proposent de la documentation aux utilisateurs de façon fluide, via des bulles d'aide pour chaque champs. Ça a l'air facile n'est-ce pas ? Oui et non.

Il est vrai que l'utilisation généralisée des standards facilite beaucoup de choses, mais comme nous l'avons déjà dit, les humains ne sont pas aussi constants que les ordinateurs, ce qui rend l'interprétation des règles parfois difficile. L'interprétation peut être différente d'un individu à l'autre, mais également d'une institution à l'autre. Pendant les quatre années durant lesquelles CENDARI et EHRI ont travaillé sur les métadonnées de plusieurs institutions en Europe et au-delà, ils n'ont jamais rencontré deux institutions qui procédaient tout à fait de la même manière ou qui suivent les standards ICA à la lettre. Il y a toujours une particularité pour toute institution. Cela signifie-t-il qu'investir du temps dans la normalisation est alors une perte de temps et de ressources ? Au contraire ! Bien que des variations soient possibles sur le thème, au moins tout le monde chante sur le même ton et le refrain est le même pour tout le monde. Employer les standards archivistiques est un premier pas important pour garantir une communication efficiente et effective de vos collections par le biais de votre propre site, mais

également pour les partager sur une plateforme commune avec d'autres institutions archivistiques et donc de stimuler leur visibilité et de s'associer à une large communauté.

Les standards utilisés le plus fréquemment, à part ceux concernant la description des collections d'archives, sont les standards décrivant les collectivités, les personnes et les familles (en jargon, les données d'autorités, voir « Norme Internationale sur les notices d'autorité utilisées pour les Archives relatives aux collectivités, aux personnes ou aux familles » ou ISAAR (CPF) et le standard ISDIAH pour décrire les institutions archivistiques mentionnées plus haut). Décrire les collectivités, personnes et familles, comment s'y prendre ? Toutes les archives proviennent soit d'une institution (comme un ministère ou autre organisation), soit d'une personne (une figure publique ou une personne privée moins connue), soit d'une famille, et les collections d'archives renseignent sur l'activité de ces mêmes acteurs. C'est pour quoi ils sont décrits comme « créateur(s) » de l'archive ou comme mot-clé (point d'accès) de la description pour expliquer de quoi il s'agit. Organiser toute l'information sur ces « autorités » est complexe. Par exemple, les organisations de l'ex-URSS sont représentées en caractères cyrilliques (leurs noms officiels), en entiers et en abréviations, ainsi qu'en anglais (ou d'autres traductions) : ainsi, la Commission de l'État, chargée de l'instruction des crimes des envahisseurs allemands-fascistes peut être dénommée par « Чрезвычайная Государственная Комиссия", "ChGK" ou "Soviet Extraordinary State Commission for the Investigation of the German-Fascist Crimes". Tout comme les organisations, une personne peut être connue sous un surnom (comme un nom de plume ou nom de scène) et on peut souhaiter enregistrer tant le nom officiel que le surnom. Autre cas, la personne décrite a changé de nom. Par exemple, vous désirez enregistrer le nom de naissance de David Ben-Gurion, David Grün. De plus, il est fréquent d'ajouter les dates de naissance ou de fondation, des informations biographiques ou l'histoire de l'institution. Encore une fois, les standards vous offrent un modèle tout fait qui permet d'organiser vos données de manière systématique. Ceci a pour avantage supplémentaire que d'autres organisations qui utilisent le même standard pourront mieux échanger les données entre elles.

Identifier de manière unique vos collections

Maintenant que nous avons défini une normalisation, il faut encore organiser nos collections et leur attribuer des identifiants. Cela semble évident, mais attendez un instant car les identifiants ne sont pas aussi simples à utiliser que l'on puisse imaginer. Nous avons rencontré beaucoup d'institutions qui changent les identifiants pour leurs collections au fil du temps. Ceci est regrettable puisque les chercheurs feront référence à vos collections avec des identifiants changeant au cours des années. Pareillement pour les institutions qui attribuent plusieurs identifiants à une description ou qui rassemblent une multitude de descriptions différentes sous un identifiant. Tandis que ceci semble logique pour la personne faisant les descriptions des collections, les relations qui ne résultent pas d'une connexion seule à seule entre les collections archivistiques et leurs descriptions, posent des problèmes pour le lecteur.

Un autre problème est le mécanisme trop complexe d'attribution des identifiants. Une institution archivistique française avec laquelle nous travaillions a commencé à utiliser une numération en chiffres romains comme identifiants pour ses collections. Alors que certaines personnes peuvent toujours lire le

latin sans hésitations, la plupart ont des problèmes pour déchiffrer la numérotation romaine, surtout quand les chiffres montent… Déchiffrer les numéros pose de toute façon problème, c'est pour cela que les utiliser n'est pas une bonne idée. Ceci est également le cas pour des numérotations compliquées à plusieurs caractères. La solution pour avoir de bons identifiants est une structure solide, simple et surtout qu'ils restent stables et persistants sans changer au fil des années. En combinaison avec le code ISIL, l'identifiant unique pour les collections est la clef pour ouvrir et partager vos données de manière durable avec le monde extérieur.

Mais est-ce que les identifiants uniques et persistants sont une solution tout-en-un pour partager vos données ? Malheureusement non : partager les données et plus spécifiquement voir où les données de votre institution se connectent aux données d'une autre institution dans un espace de partage ou sur une plate-forme commune de recherche demandera toujours un certain travail manuel pour déterminer quelles collections se connectent avec d'autres. Bien sûr, faire un travail de coréférence avec des identifiants changeant peut compliquer, voire rendre impossible le processus de coréférence. Cela vaut donc bien la peine d'investir du temps et des efforts pour produire des identifiants uniques et persistants.

Considérez quels niveaux de description vous voulez et dont vous avez besoin.

Contrairement aux armoires de classement et les livres publiés, les bases de données sont délicates dans le fait que c'est vous qui décidez, avant de commencer, comment les hiérarchies de vos descriptions doivent être organisées. Les musées ont tendance à se focaliser sur des descriptions au niveau de l'objet, puisque ceci est nécessaire pour construire des expositions, des catalogues, etc. Les centres d'archives traditionnels sont enclins à regrouper leurs collections dans des fonds, des collections ou groupes de documents et puis de décrire de façon plus détaillée à des niveaux inférieurs et dans certains cas, même au niveau de la pièce. Quand vous décidez quels niveaux de descriptions conviennent à votre institution, gardez à l'esprit que vous envisagez de partager cette information basique sur vos collections en ligne (enfin, nous l'espérons). Il faut également concentrer votre attention sur le fait que même si vous êtes prêts à partager toutes vos informations, il faut tenir compte des lois pour la protection de la vie privée et/ou les copyrights qui imposent certaines limitations concernant le partage et la publication de données autant sur votre site qu'avec d'autres projets.

Un autre point important à prendre en compte est à la capacité des systèmes de gestion et des bases de données à être flexibles pour s'adapter aux hiérarchies changeantes. Bien évidemment, il y a toujours la possibilité de créer des descriptions au plus bas niveau qui rend possible une approche descendante, mais l'inverse, une approche ascendante n'est pas nécessairement possible. Les institutions qui se focalisent sur une description du niveau de l'objet peuvent ainsi avoir des difficultés à créer des niveaux multiples entre les descriptions. De plus, ceci n'est pas faisable dans tous les systèmes employés par les institutions. Il est donc impératif de prendre en considération toutes les contraintes avant de commencer votre base de données et premièrement de réfléchir à la manière de transformer vos collections vers une structure horizontale.

Pour les archives qui préservent des copies d'autres archives, nous conseillons de prévoir une description au même niveau que celle de l'archive originale. En pratique, nous remarquons souvent que toutes les copies d'une archives sont regroupées dans un collection-copie. Dans le meilleur cas, les différents fonds et documents dont les copies ont été prises, sont mentionnés dans les guides d'archives, mais sans qu'une cote spécifique ait été attribuée à la collection-copie. Par exemple, RG12.001 peut regrouper des copies d'une archive départementale en France, mais le fait que le fonds A.36 soit copié dans cette collection-copie, n'a pas d'identifiant spécifique. Pour permettre de lier la collection originale et la collection copiée, fournir un identifiant unique aux fonds copiés facilitera la tâche. Il est donc pertinent d'y faire attention quand vous amenez des copies.

Standardiser les entrées

Comme nous l'avons déjà mentionné dans l'introduction, la standardisation de l'information dans nos descriptions est la meilleure méthode de partage. La bonne nouvelle est qu'il existe déjà certains standards internationaux que nous pouvons appliquer au sein de nos institutions. L'Organisation internationale de normalisation ou ISO a développé des codes, nommés codes ISO. Ces codes requièrent des entrées standardisées pour indiquer le pays (ISO 3166) et pour les langues (ISO 639) – il y a aussi des codes ISO pour des contenus physiques, mais ceci n'est pas nécessaire dans ce cas-ci. L'information sous forme informatisée des codes peut être employée dans la langue requise et sollicitée par ses utilisateurs. En pratique : « BE » « deu », « nld » et « fra » peuvent être lu comme « Belgium » où les langues sont « German », « Dutch » et « French » ou comme « Belgique », « allemand » « néerlandais » et « français ».

Pour l'indication des dates, qui sont dans un texte libre fréquemment notées de manière aléatoire, le code ISO 8601 – qui standardise les entrées avec le modèle [AAAA]-[MM]-[JJ] – est très utile. Évidemment, vous pouvez argumenter que vous employez déjà des formats standardisés comme « 12/11/2015 ». Ceci n'est a priori pas dérangeant, sauf qu'un européen lira 12 novembre 2015 et un américain 11 décembre 2015. Si nous adhérons à un standard, nous ne devrons plus deviner la date puisque l'affichage serait : 2015-11-12 car nous faisions référence au 12 novembre 2015 et non au 11 décembre 2015. Évidemment, si nous affichions toutes les dates intégralement, cela ne poserait pas de problème non plus. En fait, cela compliquerait les choses puisque de cette manière-là, les dates seraient limitées à une langue et ne serviraient pas l'informatique puisqu'un ordinateur ne peut pas faire le lien entre « 12 Novembre 2015 » et « 12 November 2015 » alors qu'en utilisant les standards, les deux possibilités sont regroupées : « 2015-11-12 ». En plus de ces arguments convaincants, le ICA se réfère à des codes ISO existants dans leurs directives des descriptions. Ceci renforce l'intérêt que nous devons porter ver la normalisation.

Certes, l'ISO ne propose pas une solution parfaite pour toutes les entrées que nous voudrions standardiser. Beaucoup d'institutions, plus particulièrement les plus importantes et anciennes, travaillent avec un thésaurus. À l'origine, le thésaurus était une liste alphabétique de termes standards utilisés pour le classement de la documentation, publiée sous la forme de gros livres qu'on trouvait sur les bureaux des employés décrivant les collections. Puisque une mise à jour des thésaurus n'était pas fréquente, ils restaient un outil stable. Néanmoins, de petites et grandes erreurs s'y sont faufilées au fil

du temps allant des erreurs typographiques (comme antisemitims) et des variations de certains mots (comme anti-sémitisme et antisémitisme) jusqu'à des employés capricieux qui ont inventé de nouveaux termes comme « Haine des Juifs ». À nouveau, si l'humain parvient à déchiffrer et interpréter les différences et subtilités, pour les ordinateurs, il s'agit de quatre termes complètement différents. En conséquence, un lien entre les quatre termes est impossible et il ne sera pas affiché dans les résultats de la recherche si vous recherchez un des quatre termes. Utiliser des vocabulaires standards, particulièrement avec une liste déroulante (ce qui signifie que vous pouvez choisir le mot que vous désirez à partir d'une liste), est un outil magnifique qui permet d'éviter des erreurs typographiques, de créer de la cohérence et de pouvoir partager le fruit de votre travail tant à l'intérieur qu' à l'extérieur de votre institution.

Maintenir les pièces du puzzle

Avec toutes les possibilités du monde numérique, beaucoup de sources analogiques ont été numérisées et sont préservées aussi bien dans leur forme analogique que numériquement dans les archives. De plus en plus de ressources existent uniquement en forme numérique sans forme analogique de l'original (un catalogue imprimé des ressources digitales peut évidemment exister). La préservation numérique est un défi en soi puisque cela requiert un rangement astucieux et une identification des ressources, notamment quand elles existent en différents formats. Vous pouvez scanner un document, puis utiliser l'OCR pour effectuer une opération de nettoyage manuel et identifier les entités nommées (l'OCR, reconnaissance optique de caractères, convertit des images de texte en des textes encodés et éditables ; la reconnaissance d'entités nommées identifie certains éléments dans ces textes comme les noms de personnes, d'organisations, de lieux, etc.). Cela conduit à l'apparition de quatre fichiers différents, tous provenant d'une source physique, qui émergent pendant les quatre différentes étapes du processus de numérisation. Comment alors maintenir la provenance de ces sources ? Il est important de créer une connexion, un lien entre les données numériques et leurs descriptions. Lier les données avec leurs métadonnées (donc la source avec sa description) peut se faire de plusieurs manières. Il faut à nouveau tenir compte de la manière dont votre outil de gestion des données réagira (ou non) pour pouvoir sélectionner quelles parties de vos (méta)données sont seulement dédiées à un usage en interne (uniquement pour le personnel) ou dans la salle de lecture, quelles (méta)données peuvent être partagées sur votre site web et lesquelles seront communicables avec des infrastructures de recherche, des catalogues en ligne et autres en dehors de votre institution.

Publier des informations de vos collections sur votre site web

La façon la plus simple de donner accès à vos collections en ligne est de les intégrer à votre site web. La plupart des institutions offrent un accès plus ou moins élaboré aux guides d'archives. Il peut s'agir d'une visualisation en temps réel du statut actuel de votre base de données (des mises à jour instantanées) ou peut être une sortie de votre système interne à une certaine date ce qui demande de faire des mises à jour au fur et à mesure. Les institutions peuvent choisir de restreindre ces informations à des pages d'internet dites classiques (ou des « pages html »), mais elles peuvent aussi bien décider de publier les descriptions de ses collections dans un format qui permet aux autres de les utiliser, par exemple en

utilisant .xml (extensible Markup Language, un langage qui définit des règles pour encoder un documents dans un format qui permet autant aux humains qu'aux ordinateurs de le lire) et EAD (Encoded Archival Description, un standard pour l'encodage d'instruments de recherche d'archives pour l'usage en ligne). Les abréviations semblent anodines, mais c'est l'idée derrière qui compte.

« Dans notre famille, nous partageons » - l'exportation de l'information de votre institution vers le monde extérieur.

Quand mon mari était encore enfant – il est le second de cinq enfants – sa famille fut invitée chez leur grands-parents. L'aîné, qui était très excité, finissait son assiette à toute allure, puis reluquait celle de sa grand-mère, qui était encore intacte – elle allait commencer à manger. Il la contempla pendant un moment, puis annonça tout naturellement : « dans notre famille, nous partageons ». Évidemment, cette histoire fut inscrite dans l'histoire familiale, souvent reprise et réitérée pendant les fêtes de famille. En prenant la direction d'un groupe international conçu de plusieurs nationalités et une multitude de langues et de contextes culturels, j'ai voulu créer un esprit de communauté et racontai cette anecdote pour démontrer qu'elle est parfaite pour décrire notre mission. Depuis lors, tout le monde dans le groupe s'est approprié la phrase : « dans notre famille, nous partageons ». Ceci promut une communication confortable et une atmosphère de partage positive, non seulement pendant les repas au cours de nos réunions, mais aussi dans l'espace virtuel de travail dans lequel nous travaillons. Dans l'univers numérique, il est impératif de se faire voir et d'accroître sa notoriété. Quelle institution ne tient pas de statistiques sur le nombre de visiteurs de leur site web, le nombre de « likes » qu'elle reçoit sur Facebook et le nombre de mentions sur Twitter et les retweets qu'elle génère ? Cependant, est-ce que cela nous aidera à nous focaliser sur notre propre présence en ligne ? Je ne pense pas. Je pense fermement que partager nous apportera plus d'avantages (sans compter un sentiment de joie) pour tous les participants. En fin de compte, l'effort de partager et la diminution de l'information exclusive à l'institution qui est uniquement accessible sur votre site, apporteront plus de visibilité à votre institution, ses collections et le travail assidu que vous avez investi dans la préservation et l'ouverture de votre héritage culturel important. Et n'oublions pas que l'inclusion dans une communauté peut bénéficier à tout le monde.

Toutes les étapes discutées plus haut faciliteront désormais l'exportation de l'information. Un facteur important que nous n'avons pas encore abordé est de savoir si votre système est à la hauteur pour l'exportation de données. Pour cela, voici un conseil primordial : chaque fois que vous devez faire appel à une firme d'informatique extérieure pour commencer ou pour améliorer votre programme de gestion de données, ne vous concentrez pas uniquement sur l'entrée et l'importation de données, mais renseignez-vous aussi depuis le début, dès l'incorporation des premières descriptions, comment vous pouvez sortir les données ! La maniabilité de votre système de gestion des données est basée sur la capacité de votre système d'exporter les données et dans quels formats elles sont supportées par votre système. Il arrive à maintes reprises que des archives aient un système au sein de leur institution qui ne permet aucune modification et qui les obligent donc à solliciter une firme extérieure pour exporter leurs données de leur propre système. Personne n'a envie d'être dans cette situation (et si vous êtes dans le cas, il est temps d'y remédier, ou d'adopter une attitude plus affirmée – il n'est jamais trop tard). Un

outil tel que OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) est sans doute quelque chose à considérer.

Outre la capacité d'exporter vos entrées, vous pouvez également souhaiter faire une sélection dans vos exportations, non seulement en fonction des sujets mais également au niveau de l'accès. Vous pouvez ainsi souhaiter n'exporter que les descriptions de votre collection photos liés à la Première Guerre Mondiale, ou seulement les niveaux supérieurs des descriptions de vos collections sur les procès suite à la Deuxième Guerre Mondiale, parce que les niveaux inférieurs doivent rester au niveau interne pour des raisons liés à la loi pour la protection de la vie privée. En même temps, vous souhaiteriez certainement garder les informations administratives et d'autres champs de descriptions uniquement pour l'usage interne, tout comme le travail en cours qui n'est pas encore prêt à être partagé. Idéalement, toutes ces exigences sont gardées en tête quand vous débattez du système de gestion de données qui conviendrait le mieux à votre institution. Plus vous pouvez tester de choses pendant la période d'installation ou d'essai d'un nouveau programme, mieux cela vaut.

Quand vous partagez vos informations avec le monde extérieur, il y a des chances qu'il existe des informations parallèles venant de différents systèmes. Il peut y avoir par exemple un guide de recherche qui recouvre toutes les sources concernant la Première Guerre Mondiale dans un pays, compilé par une institution centrale qui participe au même projet que vous. Les descriptions des sources de votre institution seront donc accompagnées par une description parallèle du guide de recherche. Ou alors, certains chercheurs ont pu indiquer les sources disponibles dans votre institution et décrire quelles collections étaient pertinentes pour leur propre sujet de recherche. En somme, il y a plusieurs scénarios possibles. Ceci devrait nous sensibiliser envers l'importance de notre responsabilité pour décrire clairement les sources et les références. Ainsi, il faut donc veiller à ce que la provenance de vos données soit clairement indiquée avant de partager vos métadonnées tout en incluant des informations concernant les versions, qui doivent au moins inclure la date à laquelle les dernières modifications ont eu lieu. Cela facilite la marche arrière en évitant un processus long et difficile pour retracer ces traces.

En dernier lieu, il est probable que partager vos données dans le monde numérique peut apporter des contributions à votre institution. Généralement, les environnements numériques de recherche invitent leurs visiteurs à être actifs en proposant des annotations, enrichissements, en bref : des commentaires sur l'information présentée sur le site. Les annotations des chercheurs peuvent être très utiles pour votre institution et il est même possible que vous vouliez les incorporer dans votre système. Inutile de préciser que ceci est une excellente occasion, sans rentrer dans les détails concernant les exigences techniques et le fait qu'il faut bien préciser la provenance et bien la documenter, pour instrumentaliser les ordinateurs comme support pour créer des échanges bilatéraux d'information, non seulement comme sortie pour les chercheurs mais aussi comme entrée, et donc l'enrichissement des données, par les chercheurs.

Mauvaises nouvelles pour les anarchistes : l'importance de directives et règles documentées et bien implémentées

Parallèlement à la normalisation et au partage, il est nécessaire d'adopter et de respecter des directives claires, une uniformité des processus et de la responsabilité ou la traçabilité des entrées. Tant que vous travailliez en isolement complet et expliquiez à vos visiteurs comment vous organisiez et décriviez les sources que vous préservez, vous n'aviez pas besoin d'uniformité des règles. Dans les institutions plus grandes et dans un environnement où le partage est capital, il est impossible de fonctionner sans directives documentées, des règles et sans contrôle qualité sur les entrées qui sont générées.

Voyons un exemple illustrant ce processus en considérant une institution, qui ne fut d'abord qu'un petit projet dans les années 1990 rassemblant trois personnes pour créer une exposition et collecter des matériaux pour un musée qui ouvrit ses portes en 1995. Au début, seulement une personne de l'équipe décrivait toutes les sources identifiées pour la création de l'exposition (ainsi que les renseignements contextuels). Comme l'information et le matériel s'accumulaient, elle a ressenti le besoin d'organiser l'information de manière structurée et commença donc à travailler dans une base de données Access, décrivant pièce par pièces les différents objets rassemblés. Le petit musée s'ouvrit et devint un grand succès, dépassant contre toute attente le nombre de visiteurs initialement prévu. Le musée développa d'autres formes d'activités en ouvrant un centre de documentation. De même, les fonds et collections s'accroissaient sur une courte période. Conscient des bénéfices que la numérisation pourrait leur apporter, et avec comme objectif de rassembler tous les matériaux du musée sur le thème, même si ceux-ci étaient le plus souvent des copies, le petit musée et son centre de documentation investirent une somme considérable pour la numérisation des matériaux. Les données – qu'elles soient originales ou copiées – étaient bien préservées. En revanche, fournir des métadonnées adéquates, des outils pour trouver le bon matériel et pour ouvrir ceux-ci aux chercheurs n'était pas la priorité. Cela résulta en un texte plat sur le site web de l'institution qui ne reflétait qu'une partie des riches collections, mais qui changeait aussi les identifiants de référence de ce qui était décrit quand les textes du site web étaient mis à jour. Ce n'est que ces dernières années, quand les infrastructures de recherche internationales ont souhaité partager les informations sur les collections de l'institution, qu'a eu lieu une sensibilisation et une prise de conscience pour travailler en environnement numérique et standardisé pour organiser les données aussi bien que les métadonnées. Passer à une structure standardisée de métadonnées numériques est non seulement un atout pour la visibilité de votre institution, mais aussi pour documenter l'information dans l'organisation elle-même, qui sinon resterait implicite et disséminée dans la tête de plusieurs personnes (qui ne restent pas indéfiniment dans la même position ou au même poste). En ayant un système organisé, votre institution est forcée d'établir des processus internes d'informations structurées, cela vous aide à partager vos informations via la publication des (méta-)données sur votre site, vous permet de partager avec les projets avec qui vous avez choisi de coopérer (les infrastructures de recherche), crée l'opportunité de développer de nouvelles méthodologies et permet donc de vous engager vers la publication numérique de vos catalogues d'archives de manière durable.

Aller plus loin

Si ce texte vous a convaincu de certains avantages que le partage peut vous apporter – avec la normalisation comme prérequis, nous avons le plaisir de vous offrir quelques suggestions de lectures pour aller plus loin. Bonne lecture !

General International standard Archival Description / « norme générale et internationale de description archivistique » (ISAD(G))

Cendari White Book of Archives

Collaborative EuropeaN Digital Archival Research Infrastructure (CENDARI)

Data Management Planning

Digital Research Infrastructure for the Arts and Humanities (DARIAH)

Encoded Archival Description (EAD)

Encoded Archival Guide (EAG)

European Holocaust Research Infrastructure (EHRI)

Europeana

Free your metadata

ICA-AtoM (International Council on Archives – Access to Memory database)

International Council on Archives / Conseil international des archives (ICA)

International Organization for Standardization / Organisation internationale de normalisation (ISO)

International Standard Archival Authority Record for Corporate Bodies, Persons and Families / Norme Internationale sur les notices d'autorité utilisées pour les Archives relatives aux collectivités, aux personnes ou aux familles (ISAAR (CPF))

International Standard for Describing Institutions with Archival Holdings (ISDIAH)

International Standard Identifier for Libraries and Related Organisations / « Identifiant international normalisé pour les bibliothèques et les organismes apparentés ») (ISIL Code)

Open Archives Initiative Protocol for Metadata Harvesting (OAI – PMH)

Standards in APEx